34.2 A 4.2GHz 0.3mm² 256kb Dual-V_{cc} SRAM Building Block in 65nm CMOS

Muhammad Khellah, Nam Sung Kim, Jason Howard, Greg Ruhl, Murad Sunna, Yibin Ye, James Tschanz, Dinesh Somasekhar, Nitin Borkar, Fatih Hamzaoglu, Gunjan Pandya, Ali Farhang, Kevin Zhang, Vivek De

Intel, Hillsboro, OR

For conventional single-V_{CC} processors, minimum V_{CC} (V_{min}) allowed for dynamic V_{CC} and frequency (DVF) during active operation is dictated primarily by read/write margins of SRAM cells in the last level cache (LLC). V_{min} in standby modes requiring fast reactivation is set mainly by the minimum voltage required for data retention in the LLC rather than soft error rate, since error detection and correction are used in LLCs. Single-V_{CC} processors have to use relatively large SRAM cells in the LLC to achieve the low V_{min} values needed to meet energy-efficiency goals. As a result, die area increases or the amount of on-die LLC reduces, thus adversely impacting cost/performance.

A dual-V_{CC} 512kb SRAM macro featuring active power management with autonomous compensation of PVT variation and aging impacts is designed and fabricated (Fig. 34.2.7) in a 65nm CMOS technology. This design enables high-density SRAM 136Mb/cm², for 64 to 256Mb LLC arrays in dual-V_{CC} processors while providing low active/standby V_{min} of 0.7/0.6V (Fig. 34.2.1). In contrast, a conventional single-V_{CC} design in the same technology can achieve only 95Mb/cm² LLC array density at 0.7/0.6V V_{min}, or only 1.1/1V V_{min} at 136Mb/cm² density. A 256kb SRAM block consisting of four 64kb sub-arrays, is optimally partitioned into two different voltage domains. The fixed high-V_{CC} region (V_{LLC}), operating at the reliability-limited maximum V_{CC} (V_{max}) of 1.2V, allows usage of the smallest possible SRAM cell whose active V_{min} is close to V_{max}, thus maximizing the bit density. The variable core-V_{CC} region (V_{CORE}) that shares V_{CC} with the rest of the processor core, designed for ultra-low-voltage operation, uses DVF across the 0.7 to 1.2V active V_{CC} and 0.6V standby V_{CC} to achieve the best energy efficiency while minimizing performance impacts. Minimal usage of explicit low-to-high V_{CC} level shifters, enabled by the optimal dual-V_{CC} partitioning, and use of dc power free embedded level converters in the WL and write drivers help minimize impacts on area, delay, and power (Fig. 34.2.2). This design is simpler than previously reported dual-V_{CC} SRAMs [1] and it provides an optimal balance between array density and energy efficiency. In addition, it incurs lower overheads in area, power, delay, metal resources, and V_{CC} distribution, especially compared to techniques that use row- or column-based cell V_{CC} switching to improve SRAM read/write margins.

Dynamic sleep transistors, active virtual ground clamps, and dynamically programmable reference (V_{REF}) voltages (Fig. 34.2.3) minimize active and standby power of the LLC under all conditions during the lifetime of the processor by autonomously and continuously tracking and compensating for changes in transistor characteristics, leakage currents, and V_{min} values induced by within-die (WID) and die-to-die (D2D) PVT variations and aging. Virtual-ground control in this design is more accurate than passive gated-MOS diode clamps [2] or active replica cell bias clamps [3] across extremes of PVT variations and aging. While programmable bias transistors [4] can compensate for P variation impacts to some extent, they incur significant silicon calibration overheads, and are not as effective against V and T variations. In addition, since no additional bias transistors are used for clamping, the proposed active clamping has lower area and power overheads. Instead, gate bias of a portion of the sleep device itself is controlled to provide clamping.

Clock gating and programmable deactivation times, derived from benchmark access patterns or leakage currents, are used to further reduce the LLC power. Setting V_{LLC}-V_{REF} close to zero minimizes sub-array leakage power when data retention is not required or a faulty sub-array needs to be disabled. Interfaces between active and inactive regions are designed to eliminate dc power consumption in circuits at active region boundaries. An early wake-up scheme is used where, based on block request, portions of sleep transistors in all sub-arrays in a block are activated, along with timers and decoders, one cycle before full activation of the selected sub-array. This helps minimize short-circuit currents in SRAM cells and ground-bounce noise due to sudden discharge of the virtual ground.

The chip supports (1) programmable weak-write-test-mode (WWTM) to measure cell stability; (2) instruction/data control that provide a range of data storage, access patterns, and read/write/idle sequences, all programmable via scan; (3) programmable sleep-transistor width to measure power-area-noise trade-offs; and (4) on-chip circuits to inject ground noises of programmable amplitudes, plus tunable supply impedances, needed to measure impacts on cell stability. Measurements are performed for multiple dies on a 12inch wafer for 0.6 to 1.4V V_{CC} and 25°C to 85°C. The dual-V_{CC} design operates at 2.3 to 4.2GHz for 0.7 to 1.2V variable V_{CORE}, 1.2V fixed V_{LLC}, and consumes 16 to 29mW active power at 85°C per 256kb block (Fig. 34.2.4). Sleep transistors reduce leakage power of idle sub-arrays by 30 to 60% at 0.8 to 1V standby V_{min}, for 2% area overhead. Leakage of sub-arrays that are disabled or do not need to retain data is reduced by 75% in a typical die.

Direct measurements of the virtual-ground voltage (V_{VSS}) across wide ranges of PVT and aging demonstrate clamping accuracy within a few mV of the V_{REF} setting (Fig. 34.2.5). In contrast, 100 to 400mV deviations in V_{VSS}, across PV corners alone, have been reported by other sleep transistor and clamping schemes [2, 3, 4]. Since the difference between active V_{CC} and standby V_{min} is around 100 to 200mV, accurate and efficient control of V_{VSS} across changes in transistor characteristics and leakage current is critical for effective LLC power reduction by sleep transistors, especially for processors in high-volume manufacturing. The proposed active-clamping technique demonstrates 14 to 24% smaller leakage power than passive bias transistor schemes [4] across large V and T variations while guaranteeing data retention. Changes in V_{min} due to WID and D2D PT variations and aging can be tracked easily by autonomous reprogramming of the V_{REF} settings at sub-array, block, or array levels, to further boost sleep-transistor effectiveness for power reduction in the LLC. Dynamic sleep transistors reduce the total active power, including power overheads of turning sleep transistors ON/OFF, by 23% for 1% activity (Fig. 34.2.6). Power/area overheads of the clamping circuit and level shifters are 4%/1%. Delay impacts of level shifters are easily absorbed in timing slacks. Finally, this SRAM macro can improve energy efficiency of a dual-V_{CC} microprocessor, containing 64Mb (256Mb) LLC, by 35% (23%), compared to a single-V_{CC} design, with minimal impact on performance or die area.

Acknowledgments:
The authors thank D. Finan & K. Ikeda, for chip implementation; M. Haycock, C. Webb and S. Borkar for encouragement and support.

References:
[1] K. Zhang et al., "A 3GHz 70Mb SRAM in 65nm CMOS Technology with Integrated Column-Based Dynamic Power Supply," *ISSCC Dig. Tech. Papers*, pp. 474-475, Feb., 2005.
[2] A. Bhavnagarwala et al., "A Pico-Joule Class, 1GHz, 32kByte X 64b DSP SRAM with Self Reverse Bias," *Symp. VLSI Circuits*, pp. 251-252, Jun., 2003.
[3] Y. Takeyama et al., "A Low Leakage SRAM Macro with Replica Cell Biasing Scheme," *Symp. VLSI Circuits*, pp. 166-167, Jun., 2005.
[4] K. Zhang et al., "SRAM Design on 65-nm CMOS Technology with Dynamic Sleep Transistor for Leakage Reduction," *IEEE J. Solid-State Circuits*, vol. 40, no. 4, pp. 895-901, Apr., 2005.

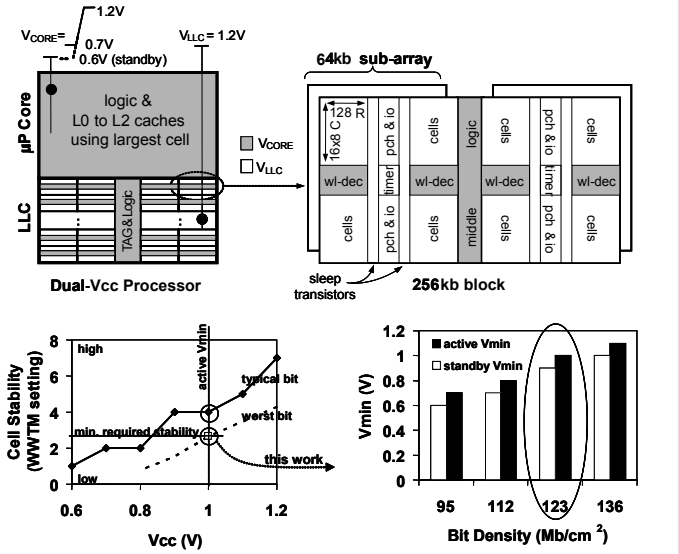


Figure 34.2.1: Single-Vcc and dual-Vcc processors.

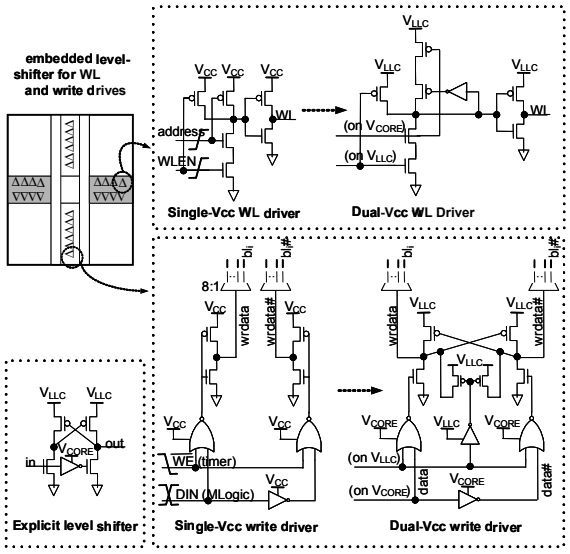


Figure 34.2.2: Comparison of explicit and embedded level shifters comparison.

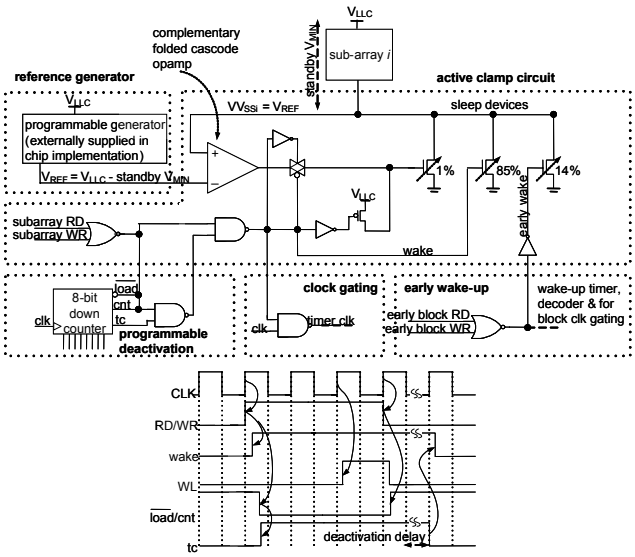


Figure 34.2.3: Power management scheme featuring autonomous clamp.

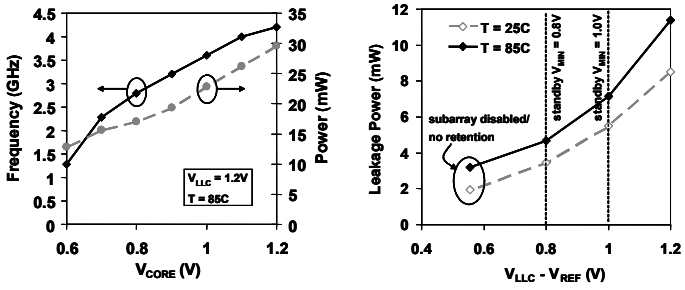


Figure 34.2.4: Measured frequency and power of 256kb SRAM block.

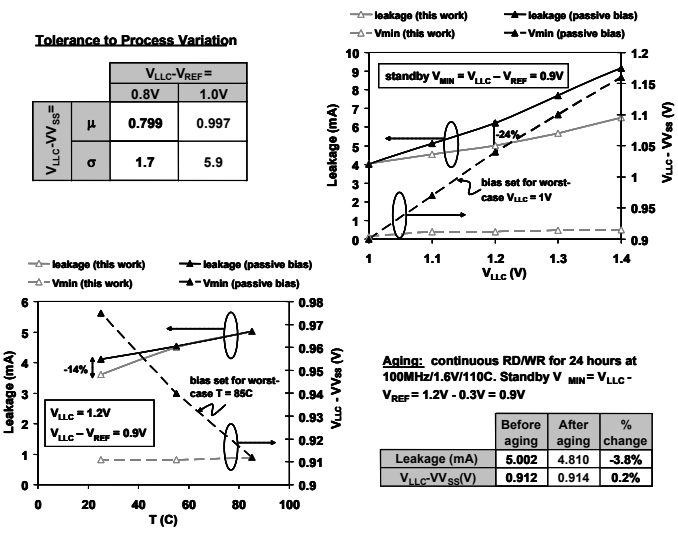
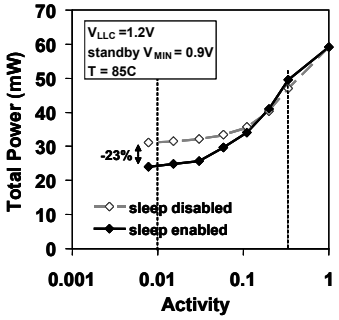


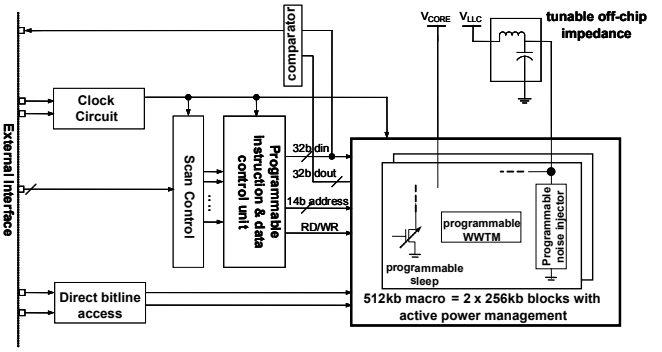
Figure 34.2.5: Measured virtual-ground voltage sensitivity to PVT and aging.



	V_{CORE} (V)	V_{LLC} (V)	Normalized Total Processor Power 64Mb	Normalized Total Processor Power 256Mb	Normalized Cache Area
Single-Vcc processor	1.0		1	1	1
Dual-Vcc with no sleep	0.7	1.0	0.74	0.91	1.002
Dual-Vcc with sleep (This work)			0.65	0.77	1.032

Figure 34.2.6: Measured power reduction and area overhead.

Continued on Page 678



Die area	1.375mm^2
SRAM cell size	0.57μm^2
Process	65nm CMOS
Interconnect	1 poly, 8 metal
Frequency	4.2GHz
Temperature	85C
Power/256kb block	30mW
Vmax	1.2V
Testing interface	membrane card
Pad count	30

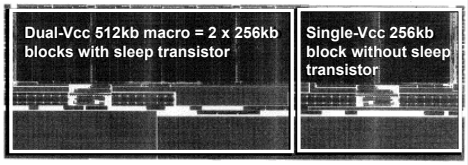


Figure 34.2.7: Chip block diagram, process characteristics, and chip micrograph.

